ISSN 1747-1524

DCC | Digital Curation Manual

Instalment on

"Archiving Web Resources"

http://www.dcc.ac.uk/resource/curation-manual/chapters/web-archiving

---

Dave Thompson
Wellcome Library
http://library.wellcome.ac.uk/

December 2008
Version 1.0

Legal Notices

Catalogue Entry

Title	DCC Digital Curation Manual Instalment on Archiving Web Resources
Creator	Dave Thompson (author)
Subject	Information Technology; Science; Technology--Philosophy; Computer Science; Digital Preservation; Digital Records; Science and the Humanities.
Description	Material on the Internet is transitory, fragile and ephemeral. Web archiving represents a systematic attempt to bring stability to the information found in websites.
Publisher	HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath; Council for the Central Laboratory of the Research Councils.
Contributor	Seamus Ross (editor)
Contributor	Michael Day (editor)
Date	December 10, 2008
Type	Text
Format	Adobe Portable Document Format v.1.3
Resource Identifier	ISSN 1747-1524
Language	English
Rights	© HATII, University of Glasgow

Citation Guidelines

Dave Thompson, (December 2008), " Archiving Web Resources", *DCC Digital Curation Manual*, S. Ross, M. Day (eds), Retrieved <date>, from http://www.dcc.ac.uk/resource/curation-manual/chapters/web-archiving

About the DCC

The JISC-funded Digital Curation Centre (DCC) provides a focus on research into digital curation expertise and best practice for the storage, management and preservation of digital information to enable its use and re-use over time. The project represents a collaboration between the University of Edinburgh, the University of Glasgow through HATII, UKOLN at the University of Bath, and the Council of the Central Laboratory of the Research Councils (CCLRC). The DCC relies heavily on active participation and feedback from all stakeholder communities. For more information, please visit www.dcc.ac.uk. The DCC is not itself a data repository, nor does it attempt to impose policies and practices of one branch of scholarship upon another. Rather, based on insight from a vibrant research programme that addresses wider issues of data curation and long-term preservation, it will develop and offer programmes of outreach and practical services to assist those who face digital curation challenges. It also seeks to complement and contribute towards the efforts of related organisations, rather than duplicate services.

DCC - Digital Curation Manual

Editors

Seamus Ross
Director, HATII, University of Glasgow (UK)
Michael Day
Research Officer, UKOLN, University of Bath (UK)

Peer Review Board

Preface

The Digital Curation Centre (DCC) develops and shares expertise in digital curation and makes accessible best practices in the creation, management, and preservation of digital information to enable its use and re-use over time. Among its key objectives is the development and maintenance of a world-class digital curation manual. The *DCC Digital Curation Manual* is a community-driven resource—from the selection of topics for inclusion through to peer review. The Manual is accessible from the DCC web site (http://www.dcc.ac.uk/resource/curation-manual).

Each of the sections of the *DCC Digital Curation Manual* has been designed for use in conjunction with *DCC Briefing Papers*. The briefing papers offer a high-level introduction to a specific topic; they are intended for use by senior managers. The *DCC Digital Curation Manual* instalments provide detailed and practical information aimed at digital curation practitioners. They are designed to assist data creators, curators and re-users to better understand and address the challenges they face and to fulfil the roles they play in creating, managing, and preserving digital information over time. Each instalment will place the topic on which it is focused in the context of digital curation by providing an introduction to the subject, case studies, and guidelines for best practice(s). A full list of areas that the curation manual aims to cover can be found at the DCC web site (http://www.dcc.ac.uk/resource/curation-manual/chapters). To ensure that this manual reflects new developments, discoveries, and emerging practices authors will have a chance to update their contributions annually. Initially, we anticipate that the manual will be composed of forty instalments, but as new topics emerge and older topics require more detailed coverage more might be added to the work.

To ensure that the Manual is of the highest quality, the DCC has assembled a peer review panel including a wide range of international experts in the field of digital curation to review each of its instalments and to identify newer areas that should be covered. The current membership of the Peer Review Panel is provided at the beginning of this document.

The DCC actively seeks suggestions for new topics and suggestions or feedback on completed Curation Manual instalments. Both may be sent to the editors of the *DCC Digital Curation Manual* at curation.manual@dcc.ac.uk.

Seamus Ross & Michael Day.
18 April 2005

Table of contents

Introduction to web archiving and digital curation 7
The nature of the Internet 7
An overview of digital curation 8
The imperatives for web archiving 9
Web archiving and digital curation 10
Acquisition 10
Curating captured Internet sites 11
Web archiving background and developments to date 12
How web archiving applies to digital curation 14
The role of preservation planning 14
Digital curation, web archiving and rights issues 14
Data creation, data creators and digital curation 15
Curating websites with the Reference Model for an Open Archival Information System (OAIS) and TDR 16
Relationships within websites 19
Data re-use, data re-users and digital curation 19
Information about curation tools 20
Web archiving in action 21
Collaboration in web archiving 21
Other issues around web archiving 22
Selection of websites 23
Curation issues and the practice of web archiving 23
Costs of web archiving and curation 24
Tools of the trade 24
Next steps 25
Conclusion 27
Bibliography 28
Terminology 30
Annotated list of key external resources 30

Biography of the author

Dave Thompson has worked in information management in both the UK and in New Zealand since the late 1980s. He's been involved in library system management and the development of on-line services and products.

Before coming to the Wellcome Library in 2006 he worked for the National Library of New Zealand on national digital preservation strategies, including projects such as the NLNZ Preservation Metadata Schema and the Metadata Extract Tool.

Dave is currently Digital Curator at the Wellcome Library in London. He leads a project that will see the Library implement its strategic aim of collecting born digital archival materials and preserving them for the long term. This work applies professional archival practice, and uses a collaborative approach to develop common tools and workflows for the acquisition and long-term management of born digital materials.

The context of digital curation

Introduction to web archiving and digital curation

In just a few years the Internet has become an important medium for personal and scientific communication, publishing, e-commerce, and much else. Increasingly it has become 'invisible' as we rely more on the services it supports and lose sight of the underlying infrastructure. The 'fluid' and unregulated nature of the Internet means that entire sites change or disappear and content vanishes, often without leaving any trace[1] and without announcing their departure. This chapter looks at web archiving and digital curation.

Material on the Internet is transitory, fragile and ephemeral. Web archiving represents a systematic attempt to bring stability to the information found in websites. To date activity has centred on acquisition and storage and not curation. The 'proper' curation of material archived from the Internet is still in its infancy.

Digital curation begins with a belief in the continued utility of material beyond an immediate period. The challenge is to find strategies and methodologies appropriate for the material being curated, but also gain the confidence of a designated user community that the adopted strategy is a valid one that will ensure them long-term access to the curated material.

The nature of the Internet

The Internet is becoming more invisible as it supports and delivers access to the services we use in our daily lives, such as Internet banking, e-mail, shopping, mobile Internet, and social web sites. The underlying technologies are increasingly geared towards ease of use and delivery, and towards facilitating a shared personal experience, recently shifting subtly from 'publishing' to 'communication'. The idea that the content we now share may have future value has not yet been embodied into the structure of the Internet.

There has been a rapid evolution of interconnecting technologies whose combined functioning now delivers Internet content. They form a complex system in which interdependence plays a key role, an environment in which almost any format of material can be made accessible. The structure of the Internet, its infrastructure, hardware, and supporting applications, are in a state of perpetual change, a state some call 'perpetual beta'. There is no standard model for the provision of Internet content. The infrastructure varies in capacity and technology between the industrial 'have' nations and the poorer 'have nots'. It is almost impossible to identify what technologies, infrastructure and supporting applications are being used to deliver any given website. Yet these may have a direct effect on the susceptibility of a website to an archiving effort.

Since web archiving cannot be separated from the environment in which its source material operates, changes in the environment necessitate changes in archiving practice. It is useful to understand what we mean by a website. Koerbin observes that, 'Since the web is an open medium for disseminating information there are obviously many more types of information or data in more

[1] Day, M., 2003, *Collecting and preserving the World Wide Web, feasibility study undertaken for the JISC and the Wellcome Trust*, http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf [Accessed; 6 Oct 2008]

formats and media that are publicly available over the web than we have had to deal with in the print publishing world'[2].

Forms of Internet publishing range from the traditionally linear—single documents containing 'pages' and 'chapters'—to more complex database driven experiences in which the next 'page' is conditional on the last and is not created until the previous has been viewed or used. The Internet increasingly supports live and highly dynamic interactions between individuals.

Whilst information and documents produced by organisations may be seen to have more formal structure, the Internet has seen an explosion of informal communication, especially blogs, wikis and on-line picture or video sites. In these the means of publication become less formal, often using stream of consciousness. A key problem in the acquisition of such large sites lies in checking that all required content has been acquired and all significant functionality retained; an almost impossible task to do manually, and a challenge that cannot be met until automated ways of checking and recording the functionality and significant properties contents of archives have been developed.

An overview of digital curation

It is a quirk of digital material that it cannot be 'put on a shelf' and left to its own devices, a simple act that has served to preserve physical items, sometimes for millennia. Day observes the Internet '…is a very difficult object to collect and preserve.'[3]

Digital curation aims to provide for the active management of born digital material. It can be the management of current or newly created material or the management of material created in the past. The term 'long-term' itself is subject to different definitions. For one designated user community data may have utility only as it is created, for another community that utility may stretch centuries into the future. The datasets created by one activity may be replaced if lost or corrupted; for another, data may be impossible to re-create at any cost.

Not all digital material should be treated identically. Digital curation comprises different sets of management processes for the different types of material and the different needs of designated user communities. What is required for the long-term management of a petabyte of astronomical data may not be what is required for the management of a terabyte of material in a web archive.

What all digital curation shares in common is that the sooner a formal curation regime begins, the more likely material will have a secure and useful life. Intervention before data is created can anticipate problems, and make management recommendations that minimise future curation difficulties, though this may represent an ideal situation. Commonly the web archivist is taking material after its release onto the Internet. If material should be archived before it is published, or if it would be more effectively archived another way then it is no longer a web archiving exercise. For some organisations

[2] Koerbin, P., 2005, *Current issues in web archiving in Australia*. Staff papers, National Library of Australia. http://www.nla.gov.au/nla/staffpaper/2005/koerbin1.html [Accessed: 6 Oct 2008]

[3] ibid

publishing on their Internet site and then backing up that site is considered—incorrectly—to be a sufficient records-keeping function.

In order to properly curate, and therefore preserve, material the curator must be in a position to exercise all necessary management control over that material. This may involve making a number of copies of the material or migrating it to a different format. It may also mean transferring the material from one hardware or software environment to another. This becomes a 'rights' matter, of the curator negotiating permissions to make such management changes in the material.

The technological infrastructure of the digital world changes constantly. Hardware becomes obsolete whilst operating systems are subject to change and replacement.

The environment in which digital material is created needs to be recorded if the life cycle of the material is to be managed. It is necessary to create and hold appropriate administrative metadata about the material that describes its significant properties, the context of the material, its origins and the history of curation activity. It is the responsibility of digital curators to collect and manage metadata for material in their care, though they may call on the knowledge held by data creators to assist them. Increasingly metadata frameworks such as the Preservation Metadata: Implementation Strategies Version 2[4] (PREMIS) provides a standardised basis for identifying metadata essential to successful long term management.

This is a more complex task for the web archivist since the many file types and formats found on the Internet each require their own metadata. Not every curator has the opportunity to place constraints on the format or types of files they are archiving and therefore are required to manage. The web archivist—increasingly the whole-of-domain archivist—may have no relationship with the creators of the websites they curate.

Rights, technological and management issues affect the process of digital curation. Rights become increasingly complex issues for 'social' websites. Where websites comprise thousands of contributions, each having the rights of its creators, addressing rights issues presents significant challenges. When all of these issues have been addressed, and only then, can full responsibility be taken for material and the necessary steps taken to achieve the goals of long-term viability and access.

Given the cost and effort involved in managing born-digital material, the assumption of responsibility for its long-term curation should not be taken lightly.

The imperatives for web archiving

There are three imperatives for web archiving:

Firstly, there is a desire to capture, at a national level, cultural or historical material or the material output of a nation. This is often expressed in the activities of a national library or national archive, increasingly turning to legal deposit legislation as a mandate for web archiving, though legislation is only slowly catching up with the social use of the Internet as a

[4] PREMIS, http://www.loc.gov/standards/premis [Accessed 21 October 2008]

medium for the exchange of thoughts and ideas.

Secondly, legal imperatives are demanding that organisations better maintain all records of their own business activities. This encompasses a broader set of activities than just web archiving, but the more the Internet is used as a publishing medium the stronger the incentive for retention. Likewise as a population relies more on the Internet for official, and especially government, communication so organisations are required to ensure that material remains available. As strong as this imperative may be, web archiving is not a replacement for good record keeping. The desire to retain evidence of Internet activity or presence may be driven by regulatory compliance but its purpose is protection against legal action and to support intellectual property claims.

Thirdly, the dominance of the Internet as a publishing medium means that cultural institutions that don't have a mandate to collect any material are using web archiving as a means of developing their collections and as a basis for additional services. Material is only being published on the Internet and only web archiving can capture this. Increasingly Web 2.0 technologies allow user communities to become more engaged with institutions.

Archiving activity to date has been driven by national libraries, archives and the **Internet Archive**. The **National Library of Australia** (NLA) is a key player and has been selectively archiving Internet-based material since 1996 for its Pandora Archive[5]. The Library undertook its first archive of the whole Australian Internet domain in 2005.

The private sector is recognising the value of web archiving. As scientific and scholarly communication utilise more shared environments such as blogs and wikis it becomes more important to record those communications. Legal compliance is not the only imperative. The rights to an individual's intellectual activity may be vested in the organisation employing them. Today's lab notes may carry economic value in the future, but only if evidence of ownership or knowledge can be shown.

Web archiving and digital curation

Acquisition is not the same as preservation and preservation is not the same as digital curation. Preservation is an aim of curation. As an activity, web archiving is mis-named. Curation is the management of many processes; all of which may result in preservation being achieved but which by themselves may not achieve this goal.

There is no single simple task that is web archiving. There are two clearly defined steps: firstly acquisition, and secondly curation. Each step is distinct from the other and may be undertaken by separate actors. The tasks may even be separated in time and space. The former task is more clearly understood and more often practised than the latter, though both may be confused with 'preservation'.

Acquisition

The process of acquisition may be highly automated or highly manual, it may be done within an archive or outsourced. The acquisition of websites continues to require knowledge of the techniques of their capture. Frustratingly every site is different. Whilst the acquisition of web pages does not necessarily equate to their

[5] Pandora, http://pandora.nla.gov.au [Accessed: 6 Oct 2008]

preservation, any action taken at the time of acquisition can have a positive or negative effect on the curation process and thus the long-term viability of the material. Often content missing from a highly dynamic web site cannot be replaced even shortly after the initial archiving activity.

Websites are acquired at the time they are most available, the majority of the file formats current and have readily available rendering applications. This relationship between data, applications and environment decays over time as obsolescence takes effect. Continued use of the archived data depends upon the maintenance of associated and supporting technologies in order to continue to have access to content.

A website's complex hierarchical structure can be archived and made accessible because the concept of hypertext mark-up language (HTML) and hypertext linking has not yet been replaced by any other mechanism, though dynamic linking though tools such as Java is being increasingly used. Likewise, the relationship between the HTML language and the hypertext transfer protocol (HTTP) also remains intact. If the fundamental architecture of the Internet were to change every website currently archived would immediately become inaccessible.

The finite process of acquisition is the least expensive, and a one-off cost, when compared to the unknown cost of on-going, long-term curation. The cost of indefinite curation can only be projected based on present costs and present technologies, though recent work such as the Life Project[6] has attempted to model future preservation costs by using web archiving as a worked example.

Curating captured Internet sites

The second step of web archiving is the means by which the interventions necessary to maintain the viability of material are managed and applied over time. It is this process of curation that currently remains least understood. Over time, change of one sort or another to the original gathered object seems inevitable. The process of curation becomes increasingly the long-term care and management, not only of the material but also of the original experience of that material.

There are two challenges when curating Internet based material. The first is dealing with the complexity of the many file formats commonly used, acquiring them with a high degree of accuracy and functionality, and retaining those qualities securely into the future. The second challenge is to manage the necessary relationships that form the functionality of websites, the relationships between individual pages and elements of pages and the relationships between formats of material and their rendering applications. Increasingly relationships exist between content of different sites or servers. This content is increasingly dependent in turn on server-side support applications, which often remain out of reach of current web archiving tools. Digital curation techniques have not yet matured sufficiently to the point where the effects of maintaining these relationships has been proven and applied.

[6] Life Project, helping to bring digital preservation to LIFE - http://www.life.ac.uk/ [Accessed: 6 Oct 2008]

Web archiving background and developments to date

In total, web archiving has been undertaken in some 16 countries[7] including the USA, Australia, United Kingdom (UK), France and Sweden. Some of these collections, however, remain hidden from users, as dark archives or pilot projects reflecting the cost and complexity of not only acquisition and long-term storage, but also of providing access to the material. For some such as the **Royal Library of Sweden**[8], rights and other legal issues have not stopped archiving but have prevented public access to archived material.

Web archiving initiatives are undertaken by a diverse set of organisations such as the **Internet Archive (IA)**[9], the **Nordic Web Archive**[10], **Pandora**[11], **UK Web Archiving Coalition (UKWAC)**[12], and **Minerva**[13]. International bodies have been formed to share ideas and co-ordinate approaches, for example the **International Internet Preservation Coalition (IIPC)**[14]. International organisations have emerged around the creation and delivery of Internet content standards, for instance the **World Wide Web Consortium (W3C)**[15].

The **Internet Archive (IA),** based in San Francisco, maintains the largest Internet archive. The IA attempts to archive the entire Internet and to make it publicly accessible. It has become a leading organisation for web archiving. It also undertakes whole-of-domain and selective archiving activity on behalf of a number of other organisations including the **National Archives**[16] in the UK and the **National Library of Australia**[17].

Web archiving remains an activity in response to material being placed on the Internet, a reactive process in which websites are selected for archiving after their publication.

Web archiving differs from other forms of archiving in two key aspects. Firstly, material is retrieved from a dynamic environment and then held as a static object. Secondly, more and more websites have no actual manifestation or presence. Database driven websites produce content as it is requested by a user, present pages structured by the use of separate style sheets and expect an end user to have common rendering applications installed to support a diverse range of material formats. Providing these pages increasingly relies on the very server-side support and host infrastructure that is typically not archived. Sometimes web archivists are trying to capture something that doesn't actually exist.

[7] PADI – Web Archiving, National Library of Australia. http://www.nla.gov.au/padi/topics/92.html [Accessed: 6 Oct 2008]
[8] The Royal Library, National Library of Sweden, http://www.kb.se/english/ [Accessed: 6 Oct 2008]
[9] The Internet Archive, http://www.archive.org [Accessed: 6 Oct 2008]
[10] The Nordic Web Archive WERA, http://nwa.nb.no/ [Accessed: 6 Oct 2008]
[11] Pandora, http://pandora.nla.gov.au [Accessed: 6 Oct 2008]
[12] UK Web Archive, http://www.webarchive.org.uk [Accessed: 6 Oct 2008]
[13] Library of Congress Web Archive Minerva, http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html [Accessed: 6 Oct 2008]
[14] International Internet Preservation Coalition, http://www.netpreserve.org [Accessed: 6 Oct 2008]
[15] World Wide Web Consortium, http://www.w3c.org [Accessed: 6 Oct 2008]
[16] The National Archives, http://www.nationalarchives.gov.uk [Accessed, 6 Oct 2008]
[17] The National Library of Australia, http://www.nla.gov.au, [Accessed, 6 Oct 2008]

Much web archiving has concentrated on client-side archiving, where crawlers behave like Internet browsers requesting 'pages' from host servers. Tools such as HTTrack, originally designed as off-line browsers, have supported many web archiving initiatives. More recently international collaborations such as that of the IIPC have developed tools specifically designed for archival web archiving. These tools actively crawl the web retrieving content directed by lists of seed URLs.

This one sided process ignores server-side functionality provided by scripting and can ignore environmental host applications such as Java that support the dynamic delivery of 'pages' to users. This is to ignore many of the limitations of archiving websites, whilst client-side archiving is relatively simple and successful it runs the risk of taking material away from the context of its original environment. Only the development of tools capable of server-side archiving will address this.

The content of the 'deep web' or 'dark web' is similarly affected. Client-side archiving skims the surface of the Internet taking the most accessible content. Material that is contained within databases or in content management systems is only accessible by user searching, a process that cannot be performed by archiving tools.

Web archiving on a large scale has had to await the arrival of other supporting technologies such as cheaper mass storage and high-speed networks. It has had to wait for the development of specialist search and display tools, especially as the size of archives grows. As the underlying architecture of the Internet changes the accessibility of archived material will be threatened. The larger the archive the greater the challenge. Decisions will have to be made whether to migrate formats to new accessible formats or to emulate archaic environments.

To date web archivists have collected whole websites as discrete objects; key information along with the trivial, the wrapper along with the contents. Whilst there have been developments in the tools for web archiving, developments to present and deliver web pages have generally outstripped the ability of tools to archive that material. Exceptions to this have been the **Nordic Web Archive** tools and new search tools such as *NutchWAX*, *Nutch* with Web Archive eXtensions, based on the *Nutch*[18] search engine.

The scale of web archiving also varies. From national legally mandated deposit libraries looking to capture national cultural and intellectual output—perhaps mandated by means of legal deposit schemes—to smaller-scale selective archiving initiatives archiving material through choice.

There are three basic approaches to web archiving that have been utilised: whole of domain, selective and thematic. None represents a complete or ideal approach. Each has been driven by the goals of the work it aims to support, the requirements on agencies to do that work or by the technology available to undertake a given type of archiving.

1. Whole of domain archives such as the *Kulturaw3*[19] project at the **Royal Library in Sweden** and an increasing number of other projects are attempting to make archives of

[18] NutchWAX, http://archive-access.sourceforge.net/projects/nutch/ [Accessed: 6 Oct 2008]
[19] Kulturarw3, http://www.kb.se/kw3/ENG/ [Accessed; 9 Jan 2006]

national Internet domains. The **Internet Archive**'s aim of archiving every domain on the Internet is the highest form of this approach.

2. Selective archiving. This involves making choices about which websites should be included, or excluded from an archive. Choices are made at the level of individual websites, typically drawn from a discrete, often national, domain and may be undertaken by a national body such as a national library. The **National Library of Australia** Pandora archive is an example of a long running selective archive.

3. Thematic approaches often focus on a discrete event of national significance such as an election or natural disaster. Choices are typically made from a discrete, often national, domain though may include material from other domains. The **Library of Congress** Minerva archive[20] is an example of a thematic archive.

How web archiving applies to digital curation

The role of preservation planning

The window of opportunity for maintaining access to archived websites has proved remarkably long. This situation cannot last indefinitely and a regime of active management—curation—will become increasingly necessary. Given the variety of formats involved the provision of a preservation plan is important as the basis for the future curation process.

A preservation plan outlines in broad terms the curation regime that material in an archive will undergo. It sets out an archives commitment to the material. It may outline which of the many available standards will be adopted to support preservation.

Digital archives are expensive to own and operate; the technical infrastructure requires expert management and maintenance. A clear preservation plan serves, in part at least, as the business case justifying the provision of the necessary resources by outlining benefits of preserving the material. Secondly, a clearly demonstrated commitment to their continued existence is explicitly required from owning institutions if an archives-designated user community is to have confidence in the archive performing its functions. Publishing a preservation plan can serve this purpose by outlining what a user community can expect from an archive.

Digital curation, web archiving and rights issues

Those creating and then curating the contents of web archives may not be the holders of the intellectual or other rights in the material being archived. The easy and very public publishing medium of the Internet and the portability of Internet-based material do not make copyright or other rights issues redundant. The possession of a mandate or directive to collect and preserve Internet-based material does not circumvent rights issues; indeed it may throw such rights into sharper relief.

[20] Library of Congress Web Archive Minerva, http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html [Accessed: 6 Oct 2008]

Material from across the Internet may easily be incorporated within one or more websites. This makes it very difficult to be certain that all material presented within a discrete site is present with the full consent of the rights owner. The understanding and management of rights issues is a complex but important aspect of the curation process.

Rights issues should be addressed at the beginning of the curation process, ideally before acquisition begins. The conditions of access to this archived material may have to have been negotiated as part of the initial agreement with the rights holders. If acquisition implies the taking of responsibility to undertake long-term management of material then it also implies an equal responsibility to provide access to that material. It is essential that access is part of the rights negotiation and that this is understood to include access over a significant period of time.

The use of proprietary formats and their dependency on proprietary rendering applications creates long-term curation challenges. It may be illegal to reverse-engineer a file format in order to convert it to one more suitable to long-term curation. The negotiation of such permissions with global software companies is probably beyond the means of smaller organisations making it a responsibility of national archives, libraries or even governments.

Data creation, data creators and digital curation

The ideal relationship between data creator and data curator is one in which both parties work closely together. As the curation process begins even a close working relationship may not be sufficient to assist in the preservation of websites. Any process of migration may move all material into formats that are unfamiliar to its original creators. Changes to underlying technologies that affect functionality may even be at odds with the intentions of the original creators. It is not unrealistic to imagine an organisation approaching a web archive looking to recover its own material after some catastrophic accident. That organisation may be quite unable to use material it had created but which had subsequently been migrated to *XML* for preservation purposes.

As the scale of web archiving increases, as archiving activity moves increasingly from limited selective archiving to whole-of-domain archiving, so does the separation of creator from curator. Selective or thematic archiving can maintain a relationship with data creators; whole of domain archiving may never achieve this.

The creation of websites has become a relatively easy and automated process. Complex sites are no longer the accomplishment of large commercial organisations. Sites such as **Flickr**[21] offer users a framework into which they can easily add their own content.

The range of data creators for websites is huge:

- Individuals—blogs and Internet-based picture albums
- Groups of individuals—sports clubs, societies, and so on
- Commercial organisations—businesses small and large

[21] Flickr, http://www.flickr.com [Accessed: 6 Oct 2008]

- Support or advocacy organisations—charities or pressure groups
- Governments—departments and agencies
- Regulatory bodies—standards agencies or professional bodies

Some creators of websites may not be aware that the material they produce may have some future value. They may not want their material to be retained over time. Even if they are aware and see the value they may not have the means to hold their material in a securely managed environment indefinitely.

Some data creators, such as governments, have internally agreed guidelines for the design, style and content of websites. Such guidelines aim to standardise site construction and presentation in a way that makes information contained in them available in common formats, available to the differently abled without requiring additional rendering applications. There has been a gradual increase of standards that can be applied to Internet design from organisations such as the **World Wide Web Consortium**. However, these standards have been focussed on content creation and delivery, not necessarily on the goals of long-term preservation.

There is increasingly an environment in which some content creators, especially government and commercial sectors, are mandated to make material publicly available and to ensure that it remains available. As a result, websites often get bigger and more complex as they retain more of their content over time. This may do little to assist the long-term curation of the material.

One of the functions of curation is to provide certainty of access. By aggregating content in an archive, defining that archive's designated user community and providing sophisticated retrieval facilities, material can be more easily identified and found. In this way an archive can add value to the material it holds.

Curating websites with the Reference Model for an Open Archival Information System (OAIS) and TDR

A useful and simple definition of digital preservation has been provided by The **Research Libraries Group**[22], and cited by Jantz and Giarlo[23]:

Digital preservation is defined as the managed activities necessary:

> 1) For the long-term maintenance of a byte stream (including metadata) sufficient to reproduce a suitable facsimile of the original document and,
>
> 2) For the continued accessibility of the document contents through time and changing technology.

Digital curation can be seen as the management and provision of this set of activities. Acquisition is at the beginning of the curation process and with it the assuming of full responsibility to keep material and to maintain its long-term viability as outlined in a preservation plan.

---

[22] Research Libraries Group, 2002, *Trusted digital repositories: Attributes and responsibilities*. An RLG-OCLC Report. http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf [Accessed: 6 Oct 2008]

[23] Jantz, R, and M. Giarlo, 2006, *Digital preservation. Architecture and technology for trusted digital repositories*. D-Lib Magazine. Vol. 11 No. 6 June 2005 http://www.dlib.org/dlib/june05/jantz/06jantz.html [Accessed: 6 Oct 2008]

Whatever the chosen preservation methodology a repository adopts there are two key tools that can be applied in support of the curation process. These are the OAIS[24] functional model and the principles of the TDR[25]. Both tools offer a flexible framework within which material for preservation can be taken into a curation regime. They are not specific to preserving websites, however, and have wide and general applicability in all digital curation.

The Reference Model for an Open Archival Information System

The OAIS reference model offers an idealised workflow in which the sequential steps of ingestion, storage, management and access are mapped and their relationship to each other described.

The model has applicability to the management of all digital materials it allows, 'in principle, a wide variety of organizational arrangements, including various roles for traditional archives, in achieving this preservation.'[26] In this sense it provides a common model for the curation process that allows the curation of all digital material to be approached in a common way and from a common frame of reference for its curators.

OAIS describes each step necessary to store and manage born-digital materials as processes of ingestion, storage, management and access. By breaking down each part of the process into discrete activities each part can be fully described and responsibilities and actors assigned to those processes.

The collection of metadata about the processes of curation is essential in tracking what has been done to material over time. It is the basic information on which all-future management decisions may be based. The collection of metadata should begin with the acquisition of the material. It is this metadata that should be inseparably associated with the archived object. The ingestion process for a repository like *Fedora*[27] automates the process of creating some of this metadata, creating it as a *METS*[28] file, bound to the material it describes. Current acquisition remains focussed on acquiring only the material and not its associated metadata. There has not yet been a full exploration of the means to capture or generate metadata that will support the curation process.

A digital repository is simply a place where digital material can be stored and managed, or more accurately 'curated'. Our experience in the long-term storage and management of Internet-based material is not yet at a mature point where certainty of preservation and perpetual access can be a feature of our curation processes. The OAIS reference model provides a useful common set of processes and a language on which long-term curation can be built. It provides an aspirational model to which owners of repositories can subscribe and against which they can both map their progress

[24] Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS). http://public.ccsds.org/publications/archive/650x0b1.pdf* [Accessed: 6 Oct 2008]
[25] Research Libraries Group, *Trusted Digital Repositories: Attributes and Responsibilities An RLG-OCLC Report. http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf* [Accessed: 6 Oct 2008]
[26] Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS). http://public.ccsds.org/publications/archive/650x0b1.pdf* [Accessed: 6 Oct 2008]
[27] The Fedora digital repository, http://www.fedora.info [Accessed: 6 Oct 2008]
[28] METS: Metadata Encoding &Transmission Standard Official Web Site, http://www.loc.gov/standards/mets/ [Accessed: 6 Oct 2008]

and plan further development of their curation activity.

Trusted Digital Repositories

The Trusted Digital Repository[29] (TDR) model is one that attempts to place a framework over a repository that demonstrably supports such basic principles as authenticity, accuracy and sustainability. It describes the features of a digital repository, any repository, that achieve this. To be able to describe one's repository as being 'trusted' is a worthwhile aim.

A TDR can be defined as, 'one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future.' [30]

In an environment dominated by obsolescence and change the TDR serves as a framework within which surety and certainty can be provided. The framework is broad enough to accommodate different situations, technical infrastructures, and institutional responsibilities while providing a basis for the expectations of a TDR. It sets out to define the essential characteristics of a repository in which a user of the material can have confidence; in short, trust that what is being offered is indeed what it claims to be.

The TDR takes a two-fold approach: it addresses both technical characteristics of a repository as well as more social or organisational aspects. This makes it clear not only that a whole organisation is responsible for the curation role but also where individual responsibility lies.

The essential features of a TDR can be defined as:

- Compliance with the Reference Model for an Open Archival Information System (OAIS)
- Administrative responsibility
- Organizational viability
- Financial sustainability
- Technological and procedural suitability
- System security
- Procedural accountability

Summary of tools

The strength of the OAIS and TDR models lie in their universality. Neither imposes one management strategy over another. Neither suggests one tool over another and neither relies on specific technical solutions. Their value lies in their offering common frameworks within which many communities can work and many repositories can function. Curators of material can then communicate and share their work using common principles.

Nor are these two tools yet fully mature. Aspects of each are beginning to be incorporated into digital curation but they are not yet available as 'off the shelf' applications. The very flexibility that lends both OAIS and TDR to system, software, platform and organisation independence mean that any organisation wishing to adopt either must develop their own implementations. For all organisations

[29] Research Libraries Group, *Trusted Digital Repositories: Attributes and Responsibilities An RLG-OCLC Report. http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf* [Accessed: 6 Oct 2008]

[30] Research Libraries Group, *Trusted Digital Repositories: Attributes and Responsibilities An RLG-OCLC Report. http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf* [Accessed: 6 Oct 2008]

curating digital material, the models provide a common yardstick by which individual activity can be measured.

Relationships within websites

Within a website a complex set of relations exits between each of the components. The HTML code of a website describes and contains the necessary relationships between components; it defines the structure of both the overall site as well as individual pages. Increasingly these internal relationships are described programmatically, for instance when content is provided by a content management system.

The removal or loss of either a component—a 'page' or image—affects the integrity or functionality of the whole. As complex objects, the relationships between components must be retained as part of the curation process. Many of these mechanisms of construction are hidden from the web archivist, or only hinted at. Yet they need to be understood and maintained, to preserve an archived site's integrity.

Other relationships also apply between the non-HTML components of sites and their necessary rendering applications. Specific formats require additional rendering applications, for example, a document in *Adobe PDF* format requires a reader application to be present on a user's computer before the file can be read. Additionally, rendering applications must also be of the appropriate version to the file they are rendering.

The relationships within a website and between the components of a site and their own rendering applications forms a matrix that must be understood and captured as future curation metadata if the original object is to be understood.

These relationships create management issues from the very beginning. Some web archiving tools—such as *the Web Curator Tool*—modify the functionality of these links, both internal and external, to create an object that can function outside the environment from which it was taken. External links, that is links to websites on a different domain, can be redirected to an intervening page indicating that a website is external to the one archived rather than taking a user directly to that site.

Data re-use, data re-users and digital curation

Websites typically serve a diverse audience. A single website is used to provide access to informational outputs of an organisation and to more general information about that organisation. Information is contained within discrete publications, for example, annual reports, annual accounts or datasets. Information products or services may be offered and made available as audio or video. Whilst exceptions exist, a single site typically attempts to serve many communities.

A web archive may be created for a specific designated user community but those who use it may do so for many different reasons, from critical historical research to simply being curious. The website that acts as the wrapper for its content may itself have value to those studying how websites were constructed at a point in time. The general nature of sites makes it hard to identify any single designated user community. In this sense web archives follow the traditional model of acquisition where material is held in the

belief that it has value even though there may be no immediately identified user.

Digital material is uniquely mutable, its very portable and transformable nature makes it able to be re-used and re-presented in ways not originally thought of. That the same material can be re-worked for different designated user communities can be the benefit of the careful curation of material and the creation of rich metadata. Many websites of smaller organisations do not have a search facility. Placing that site into an archive with the ability to search those pages immediately exposes the content of those sites to scrutiny in a way that they never were originally. Archived material can be used to prove, or not prove, an organisation carried out any activity.

The basis for the re-use of archived websites is the metadata that has been collected and created for that material. Rich metadata assists curation as well as resource discovery. This can be especially valuable when dealing with the large archives that result from whole-of-domain archiving. Yet here there is a tension. In order for the granularity of retrieval to be fine enough the metadata must be equally fine-grained. In a large body of data this can only be realistically accomplished if the process of creating that metadata is automated. However useful a whole-of-domain archive is it is useless without access that matches the scale of collection.

Information about curation tools

Digital curation is a rapidly evolving activity undertaken by a global community, one willing to share and discuss aspects of their work. There are a number of tools that are available, some are stable products others still in development, and many have come out of projects that have been investigating aspects of digital curation in one form or another.

Given the dynamic nature of this work it may be meaningless to list what is currently known here. However two sources of information are worth noting. The **Digital Curation Centre**[31] maintains an extensive list of curation and preservation tools and applications on its website. The list is available at http://www.dcc.ac.uk/tools/digital-curation-tools/.

The **National Library of Australia** hosts the Preserving Access to Digital Information (PADI) web site. The list is available at http://www.nla.gov.au/padi/. This comprehensive list is a subject gateway to international digital preservation resources and supports discussion on many aspects of digital preservation. A summary of PADI discussions 'DPC/PADI What's new in digital preservation' has been jointly produced by the **Digital Preservation Coalition** and PADI. This summary is available from http://www.dcponline.org/graphics/whatsnew/.

There is also a very useful web archiving bibliography held on the website of the **Austrian On-Line Archive**. The bibliography is available at http://www.ifs.tuwien.ac.at/~aola/link/WebArchiving.html.

[31] Digital Curation Centre, http://www.dcc.ac.uk/, [Accessed: 6 Oct 2008]

Web archiving in action

Web archiving in action

Web archiving works successfully in many countries; web archiving initiatives continue to spring up and there are an increasing number of international conferences on the topic acknowledging the importance of this activity.

Archiving activity to date has concentrated on the acquisition of whole websites. It has been seen as 'web archiving' and is not yet full 'web preservation'. It is a measure of the immaturity of web archiving that much of it remains a somewhat dark art.

Collaboration in web archiving

Web archiving is well suited to collaboration between agencies that want, or need, to collect websites, or Internet-based material. The resource requirements can be shared between partners, reducing individual costs. This can be especially true where an Internet archive is being built for the first time and the high set-up costs can be shared, as was the case of the **UK Web Archiving Consortium**.

The principle of collaboration is one in which cost and effort are shared and thereby minimised whilst the opportunities for learning, and for the successful survival of archived material, are maximised. In an immature environment this can quickly produce a pool of talent and expertise. The shared learning between partners in a diverse group can call upon and further develop the skills of that group. The introduction of a single point of new expertise into a group can support the entire group by distributing new skills.

Collaboration is also an environment in which consensus can be developed within an archiving community, for instance at a national level. This can mean that consensus can be achieved around using standards-based tools such as OAIS and TDR. Use of these common open applications can support their further development as well as promote their use throughout the wider curation community. This consensual approach also supports the greater awareness of digital preservation issues within a larger community.

Successful Internet archiving requires a number of skills: technical, administrative and managerial. Not all these skills may exist in any single agency. Librarians or archivists may be equipped to fulfil the collection needs of a selective archive but less so to provide the technical skills required for a whole of domain archive, or indeed for long-term curation.

Collaboration can function well within a single organisation where the skills of librarians and archivists are partnered with the technical skills of IT staff. Increasingly IT skills are required in web archiving to not only understand the environment in which websites exist but to provide technical solutions to increasingly complex problems such as secure storage, the maintenance of a website's functionality and the development of access services.

Collaborative activity need not lower the profile of individual collaborators. Archived contributions made by individual agencies can be surfaced within their own collections as well as forming part of a collective whole. This increases opportunities for access to material, further maximising its utility and value.

For individual contributors to a consortium, providing access to material they have archived through their own catalogues also places material within the context of that collection.

Other issues around web archiving

Websites are not necessarily resistant to archiving, but their acquisition requires ingenuity and a creative mind. Web archiving is not a matter of selecting a URL, pointing an archiving tool at that URL and hitting the button marked 'archive'. That archiving should be this simple, and isn't, is another reflection of the immaturity of this activity.

Permission-secured selective archiving is costly in terms of time and effort. Whilst some sites are selected for archiving they may lack organisational information such as ownership or contact details or a postal address to which permission requests can be directed. On some sites it becomes clear that contact e-mail addresses are seldom, if ever, checked. In such cases the web archivist may have to resort to domain registries or even official lists of bodies to obtain contact details.

Some ISPs block the use of archiving tools because in the past their unrestricted use has created problems such as server overload. Some tools can place unrestrained demands on host servers by utilising maximum download speeds, maximum number of server connections and maximising transfer speeds. Whilst proper server configuration and management can minimise the impact of poorly configured archiving tools, their unrestricted use can cause servers to crash. In extreme cases this may be taken to be a denial-of-service attack by the host. In most forms of web archiving communication is with rights holders and not with hosts. How to incorporate hosts into archiving activity remains a challenge to be addressed; yet it will be increasingly essential to have both their cooperation and their support if archiving is to be successful.

Even if a site has been archiving successfully over a period of time problems can unexpectedly arise. Sites change from archivable to unarchivable as they are 'updated' or receive a makeover. What had been a small site comprising flat HTML files may re-appear as a site with Java-based navigation or functionality. The new site may be unarchivable and a gap will appear in the archive until such time as the limitations of current tools can be overcome.

It can be difficult to estimate how much space any website may occupy on storage media once it has been gathered. All sites vary in size. Some dynamic sites may vary in size from day to day. For the digital curator this can be very frustrating. It is possible to estimate the size of a website prior to acquisition but this can be inaccurate. A commonly asked question of network planners is 'How much space will you need this year for your web archive?' It can be very difficult to plan storage without past archiving activity to act as a guide. This can introduce an element of doubt and uncertainty in an organisation about the rigour of web archiving that it cannot answer such basic questions. Whilst there is no simple answer to this, future plans and estimates should be made with the close co-operation of IT and network planners to ensure that they grasp the issues and difficulties and can apply their own expertise in providing solutions.

Selection of websites

For some institutions, the demands of new legal deposit legislation are making broad selection criteria a necessity, and a mandated condition in the collection of websites. Around the world, legal deposit legislation has changed, such as in New Zealand[32], or is changing, as in the UK, to reflect the importance of the digital publishing environment. The concept of book, magazine, journal or newspaper is being replaced by the concept of 'publication'.

Curation issues and the practice of web archiving

Without permission to archive and 'hold' material, either explicitly obtained or given by legislative mandate, there can be no proper life-cycle management of that material. Rights must cover not only the acquisition of material but the ability to store and deliver material to an end user as well as the right to perform such preservation intervention as is necessary to maintain long-term viability. There are three basic approaches to rights management in web archiving:

1. Rights secured. This approach has been to secure permission to archive a site from its owners or creators, asking them to confirm that they either hold rights over the material it contains, or have cleared all rights with those that do. This has been the approach taken by the **National Library of Australia** *Pandora* archive and the **UKWAC** archive. This approach is unsatisfactory, as it is both time-consuming and expensive, and can result in less material being archived, since the implication is that unless permission to archive is granted then it cannot happen. In securing permission to archive, a closer relationship with the rights holder is obtained and this can be advantageous; not least for raising rights issues.

2. Rights assumed. This is the simplest approach and involves archiving material without explicit permission or contact with rights holders. In this model an organisation archives websites and allows rights holders to 'opt out' of the archive and request their material be removed. The Internet Archive has adopted this model. This approach is also unsatisfactory. In an archive intended to be as fully comprehensive as possible it can create gaps in the record. Any rights holder is free to ask for the removal of their material. By 'side stepping' permission, and the useful contact with rights holders, an archive may find itself faced with legal threats if material is not removed. However, if the aim is to create a fully comprehensive archive then contact with each and every rights holder or producer is an impossible goal.

3. Mandated or legislated role. National bodies under modified and updated legal deposit legislation can undertake this model. Crucially, rights are not waived or removed by legal deposit legislation. An additional layer of rights to acquire and make available material is vested in a national body such as a national library or archive. Alternately, legal responsibilities may be placed on the

[32] *The National Library of New Zealand (Te Puna Matauranga o Aotearoa) Act 2003*, http://www.natlib.govt.nz/about-us/role-vision/the-legislation-that-governs-us. [Accessed: 6 Oct 2008]

> producers of websites to place their material with such a body. This model works best in a environment where mature processes and tools exist for the long-term curation of such material since national libraries or archives have a legal duty to retain such material in-perpetuity

It would be an impossible task to undertake a permission-secured, whole-of-domain archiving exercise, as contacting or even identifying every domain or site holder would be time-consuming, hugely expensive and probably ultimately ineffective. However, it may be possible to adopt one or more approaches within the same archiving activity. For instance, rights-secured selective archiving could be used alongside rights-assumed archiving. This would allow archivists the opportunity to acquire material without first seeking permission, while giving rights holders the opportunity to remove sites for which archiving permission had not been granted.

Costs of web archiving and curation

The high cost of web archiving may reflect the immaturity of this activity, however, the cost can somewhat be offset by the flexibility and re-purposeable nature of digital material. The cost of undertaking web archiving depends upon the chosen model, the purpose of the archive and the intended period of duration.

A study by the National Library of Australia[33] found that the biggest cost for web archiving was the cost of employing people to do it.

Selective permission-secured web archiving is probably the most expensive form of archiving. It is a time-consuming activity currently undertaken by skilled, and therefore expensive, individuals. Time is spent in selection, in securing permission to archive and in managing the inevitable paper audit trails. The cost of acquisition alone can deter archiving activity from moving from project to service. Yet the cost of long-term curation can only be estimated.

There may be fewer staff involved in whole-of-domain archiving, but the costs of the infrastructure may be high if undertaken 'in-house'. Few organisations, outside national bodies, have the infrastructure to contemplate whole-of-domain archiving. There are no tools that will yet automate the capture of metadata alongside websites, and the creation of this can also be time consuming and costly.

The cost of long-term curation for this material remains uncertain. Much of it cannot be re-acquired if it becomes lost or damaged. Technical and infrastructure costs can be estimated, in terms of capital cost to buy hardware, and so on. However, as long as humans are required to have input into any curation process, the accuracy of these cost model falls off. It seems likely that as time passes and material becomes older, then increasing amounts of human decision-making and intervention may be required and the curation task becomes ever more complex.

Tools of the trade

Web archiving uses software tools to automatically download websites, based

[33] Philips, M, 2005, *Selective archiving of web resources: a study of acquisition costs at the National Library of Australia*, http://worldcat.org/arcviewer/1/OCC/2007/07/10/0000068921/viewer/file1.html [Accessed: 6 Oct 2008]

on discrete user selected URLs, or specifically selected 'seed' URLs. The URLs may be individually selected by a human, defined programmatically, or be harvested by a fully automated and autonomous application. There are few off-the-shelf curation packages for archiving websites. Archiving tools such as *HTTrack*[34] or *Wget*[35] have evolved as the Internet has evolved and have changed their functionality as changes in the Internet environment have dictated. Tools such as these are designed to acquire material and do not provide long-term curation functionality. The Web Curator Tool, developed by a consortium of organisations under the auspices of the International Internet Preservation Coalition (IIPC) continues this functional trend.

For whole-of-domain archives selection operates at a different level: that of an entire top-level domain, with the aim of complete inclusion. Archiving tools such as *Heretrix* are designed to work unattended and archive large numbers of websites. *Heretrix* has been designed to support large-scale web archiving with the intention of long-term curation, for instance in its use of the portable and open *ARC* file format. However, it does not provide an environment in which long-term curation can take place.

The PANDAS tool developed by the NLA was the first tool to provide an end-to-end web archiving application. It is based on HTTrack and offers a graphical user interface available as a web service. PANDAS stores native HTML files and it has been used successfully for a number of years by the NLA.

Under the auspices of the IIPC, the British Library and National Library of New Zealand has developed the Web Curator Tool (WCT). It is based on Heretrix and also offers a graphical user interface available as a web service. This tool offers a flexible approach to web archiving and aggregates archived material into the ARC format.

Many acquisition tools are not yet integrated with existing curation environments. Repositories such as *Fedora* or *D-Space* may have a role to play, but remain focussed on the repository/storage functions, and have not yet evolved into dedicated curation tools. The *Fedora* repository offers a useful model for a future curated archive in that it generates a metadata record at ingestion, binds that metadata to the archived material and provides a record of subsequent change to that metadata. That metadata may not be specific to the curation of websites but the principles of metadata management in *Fedora* point the direction for future development.

Next steps

The curation of archived websites is still in a period of immaturity in which acquisition and storage are the activities receiving most attention, though focus is beginning to turn to long-term curation.

The professional skills of librarians and archivists remain relevant to the curation of websites but need expanding into new technical areas. The technical skill of IT professionals and computing scientists is also relevant, but needs expanding into areas of long-term information management. Too few of either group has a sufficient understanding of the other's potential in this field; neither group has

[34] HTTrack, http://www.httrack.com [Accessed: 6 Oct 2008]
[35] GNU Wget, http://www.gnu.org/software/wget/wget.html [6 Oct 2008]

sufficient skills to fully implement a digital curation regime for websites on their own. In the future the skills of both groups need to be combined, and through collaboration successful curation regimes can be developed.

Because the environment from which websites have been taken remains largely unchanged, the curation tasks currently remain relatively simple and few. The issues around the complexity of Internet-based material are understood; the many file formats, the internal relationships between parts of a website and the relationships between those files and their rendering applications. This complexity has yet to be maintained at a point where these relationships exist in an environment in which many components are obsolete, no longer available or no longer function on current hardware.

In the future the infrastructure of the Internet as we know it will be replaced, and material gathered today may no longer function. To prepare for this future all digital material, and not just websites, should be acquired with long-term curation in mind. Current client-side web archiving skims the surface of the Internet, taking only easily accessible content, and ignoring deep web content along with host-provided software services such as Java. Server-side archiving requires the development of tools that are easily used, and are sophisticated enough to automate the capture of this part of the Internet.

The future looks to be one in which the tools of long-term curation, such as OAIS and TDR, become more refined and incorporated into everyday management applications. Without this development of a standards base, curation can never move from the current activity of copying files from server to server, to the proper life-cycle management of digital assets. OAIS and TDR are essential tools on which to base long-term curation.

The differences between curating one piece of digital material and another are not so great. The complexities of managing common formats are similar and so further development of these common tools can benefit a wide community of digital curators.

At present web archiving, especially selective archiving, remains largely a handcrafted activity. More comprehensive automation has to be a future goal. The handcrafted archive is too time consuming and costly to acquire and sustain, and so fails one of the key essential features of a TDR. Separate processes currently perform essential management tasks, such as the acquisition of metadata, or the ingestion of material into a repository. Automation is a key for the future, especially automating the process of metadata extraction.

Metadata currently plays a minor role in the archiving of websites. In an environment in which archived material can be readily accessed and complex interrelationships remain functional, the role of metadata is artificially minimised. The necessary move from acquisition-based activity to ‘proper’ curation should be built on a foundation of the sound management of metadata. Without this, the material acquired today is in danger of slipping through our fingers because we have not recorded the means to understand it. For the curation of web archives this means the development of repositories and curation processes that are built on the foundations of standards-based tools like OAIS and TDR. Repositories in which

adequate metadata is not only stored but closely integrated with the material it describes.

Conclusion

Material on the Internet has proved to be transitory, fragile and ephemeral. Web archiving represents an attempt to bring stability and permanence to information on the Internet and digital curation is the means to this end. To date activity has centred on acquisition, of copying files from server to server and not full life-cycle management—curation—of the acquired material. It is a measure of the immaturity of web archiving that much of it remains a somewhat dark art.

Internet material is increasingly being placed alongside other material held in libraries and archives, increasing the overall relevance and value of institutional collections. However, the acquisition process is complex, and can be foiled by simple technical problems, adding complexity and cost to the acquisition and eventual curation process.

There has not yet been a full exploration of the means to capture or generate metadata that will support the long-term curation of websites. Websites acquired at the present may well present significant future curation challenges, as they lack adequate metadata for their long-term management. The retrospective generation of this metadata may prove to be time consuming and expensive to undertake.

Whilst becoming more common as an activity, web archiving remains misnamed. The 'proper' curation of material archived from the Internet is still in its infancy. Practicable applications of the emerging tools are currently lacking. Standard tools form the basis for future work; TDR and OAIS are beginning to emerge as common frameworks and should be aspired to.

Rights issues continue to challenge the web archivist. The culturally interesting social networking sites, wikis and blogs, present many rights issues that are not easily overcome. In the open, free and unregulated environment of the Internet rights issues are not equally understood by all.

Despite our apparent dependence on the Internet, very little attention has been paid to the long-term preservation of websites. Whilst there is a danger that invaluable digital resources may be lost to future generations, current web archiving is too time consuming, costly and complex when done by hand. Automation, through the use of common tools, is the future. Successful curation will depend upon the generation of adequate metadata for material, and the automation and storage of metadata is a key requirement if web archiving is to become the long-term curation of websites that have portability into the future.

Bibliography

In Print

Cundiff, M., 2004, *An introduction to the Metadata Encoding Transmission Standard (METS).* Library Hi-Tech, Vol 22 No 1 p 52

Simpson, D., 2005, *Digital preservation in the regions. Sample survey of digital preparedness and needs of organisations at local and regional levels. An assessment carried out from December 2004 to March 2005,* Museums, Libraries and Archives Council

Williamson, A., 2005, *Digital Directions Strategies for managing digital content formats*, Library Review, Vol 54, No 9 p 508

On-Line

Consultative Committee for Space Data Systems, 2002, *Reference Model for an Open Archival Information System (OAIS).* http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf [Accessed: 9 Jan 2006]

Day, M., 2003, *Collecting and preserving the World Wide Web, feasibility study undertaken for the JISC and the Wellcome Trust*, http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf [Accessed; 9 Jan 2006]

Digital Curation Centre, JISC, CCLRC, British Library, 2006, *Report of the Warwick Workshop, 7-8 November 2005, Digital Curation and Preservation: Defining the research agenda for the next decade*, http://www.dcc.ac.uk/training/warwick_2005/Warwick_Workshop_report.pdf, [Accessed; 11 Jan 2006]

Koerbin, Paul. 2005, *Current issues in web archiving in Australia*, *Staff papers,* National Library of Australia. http://wwwnla.gov.au/nla/staffpaper/2005/koerbin1.html [Accessed; 1 Jan 2006]

PREMIS Editorial Committee. *PREMIS Data Dictionary for preservation Metadata version 2.* 2008. http://www.loc.gov/standards/premis [Accessed 21 October 2008]

Research Libraries Group, 2002, *Trusted digital repositories: Attributes and responsibilities*. An RLG-OCLC Report. http://wwwrlg.org/longterm/repositories .pdf [Accessed: 1 Jan 2006]

Research Libraries Group, 2005, *PREMIS (PREservation Metadata: Implementation Strategies) Working Group.* Final Report of the PREMIS Working Group. http://www.oclc.org/research/projects/pmwg/ [Accessed: 20 Oct 2008]

Research Libraries Group, 2006, *Data Dictionary for Preservation Metadata*, http://www.oclc.org/research/projects/pmwg/premis-final.pdf [Accessed: 11 Jan 2006]

Terminology

Analogue—material that has a physical manifestation as in paper or film.
Born digital—material that has been created in a digital environment and format and for which there is no analogue equivalent.
Dark Archive—an organised and structured archive of Internet based material to which there may be no public access.
Web archive—an organised and structured archive of websites or other Internet based material taken from the 'live' Internet and held as a static object with a view to maintaining viability and access for a period of time.
Whole of domain archiving—archiving activity that attempts to gather a snapshot of all Internet based material from within an entire top-level domain, such as .UK or .AU regardless of existing selection criteria.

8. Related Curation Manual chapter or other Digital Curation centre products

Appraisal and selection
Complex digital objects
Certification and trust
Ingest
Cost
Copyright and other legal restrictions
Metadata
Preservation repository models

9. Annotated list of key external resources

The British Library, the National Library of New Zealand, Sytec Resources Ltd. *Web Curator Tool*. http://webcurator.sourceforge.net/, [Accessed; 6 Oct 2008]

The Web Curator Tool (WCT) is a joint development project by the British Library, the National Library of New Zealand working under the auspices of the International Internet Preservation Coalition (IIPC). WCT is a web archiving tool, built around Heretrix, is designed to work at a national institutional level. It has been designed for managing the selective web harvesting process.

Cornell University, *Digital preservation management: implementing short-term strategies for long-term problems*, http://www.library.cornell.edu/iris/tutorial/dpm/, [Accessed: 6 Oct 2008]

This digital preservation management tutorial aims to provide a practical introduction to definitions, key concepts, practical advice and exercises in practical digital preservation. The self-paced tutorial uses a themed approach to the topic with easy-to-follow sets of questions. It is particularly geared toward librarians, archivists, curators, managers, and technical specialists.

European Archive. http://europarchive.org/ [Accessed 20 Oct 2008]

The European Archive is a non-profit foundation working towards universal access to all knowledge. It is actively building a web archive, both for itself and for other organisations. The archive also offers training courses in web archiving.

International Internet preservation Coalition (IIPC). http://netpreserve.org [Accessed 20 Oct 2008]

A coalition of 12 international cultural institutions founded the IIPC in 2003 with the aim of preserving Internet content for future generations. The IIPC has working groups on standards, harvesting and access. Their website provides access to a number of useful reports.

International Internet Preservation Coalition (IIPC) *Web Curators Mailing List*. http://netpreserve.org/about/curator.php [Accessed 20 Oct 2008]

In 2008 the IIPC established a mailing list as a discussion forum for all those interested in web archiving. The list is designed to support discussion of the practical aspects and issues of web archiving and to offer guidance and support by bringing together experts from around the world.

LIFE (Life Cycle Information for E-Literature). http://www.life.ac.uk/ [Accessed 20 Oct 2008]

The LIFE Project is a collaboration between the British Library and University College London. It has developed a model for the lifecycle of digital material that calculates the cost of 'preservation' for that material. Whist the project has a focus on E-literature it has taken web archiving as one of its cost models.

National Library of Australia, *Preserving Access to Digital Information:* PADI, http://www.nla.gov.au/padi/ [Accessed 6 Oct 2008]

The PADI website, hosted by the National Library of Australia, is a subject gateway to issues surrounding digital preservation. Like the JISC site it provides a broad discussion forum on a broad range of digital management issues. Its international advisory group provides the site with focus and relevance.

PREMIS Editorial Committee. *PREMIS Data Dictionary for preservation Metadata version 2.* 2008. http://www.loc.gov/standards/premis [Accessed 21 October 2008]

This second version of the PREMIS data dictionary provides a useful extension to earlier work. It sets out a comprehensive framework within which metadata relevant to long term life cycle management can be both identified and structured.